A tool a day keeps the bad review away

Pavol Harar[1,2] 0000-0001-5206-1794

OEAW Summer School on AI

Vienna, 13. – 17. September 2021

[1]University of Vienna, [2]Brno University of Technology

Abstract

During the life cycle of our machine learning research projects, we interact with source code, data, and models almost daily. Therefore, it is wise to make use of various tools and tricks of the community that
1) save a lot of time during the development (versioning),
2) can be decisive during the peer review (reproducibility), and
3) are invaluable for making an impact (presentation).

Moreover, the community is constantly coming up with tools useful even before the beginning (to find ideas) or after a seeming end of the project (publicity). Together, in this hands-on session, we will use those tools while developing a specific ML project. After the session, you should be able to use the tools in projects of your own.

Outline

Resources

Environment

Versioning

Presentation

Compute & Deploy

Publishing

Resources

ArXiv Sanity Preserver

- List view of the papers with abstracts and page thumbnails
- Recent ArXiv papers from `cs.[CV|CL|LG|AI|NE]/stat.ML`
- Similar papers (based on fulltext tf-idf)
- Trending papers
- Personal library
- SVM recommendations
- Open-source project by @karpathy which you can fork

Connected Papers

- Graph of related papers
- Semantic scholar database
- Visual overview of a field
- Discover the most relevant prior and derivative works
- Quick filter using titles and abstracts
- Export to BibTeX
- Collaboration with ArXivLabs

Papers with Code

Papers With Code

- Papers with available code sorted by popularity
- Openly licensed under the same license as Wikipedia
- Anyone can contribute
- Popular data sets, tasks and benchmarks
- Trending methods
- New portals also for Physics, Math, etc.

Google Dataset Search

Google Dataset Search

- Search engine for data sets
- Started in September 2018
- Indexes data sets found during crawling the web
- Quick information including License, Description, Version, etc.
- Data connected with Google Scholar
- Easy to get your own data set indexed

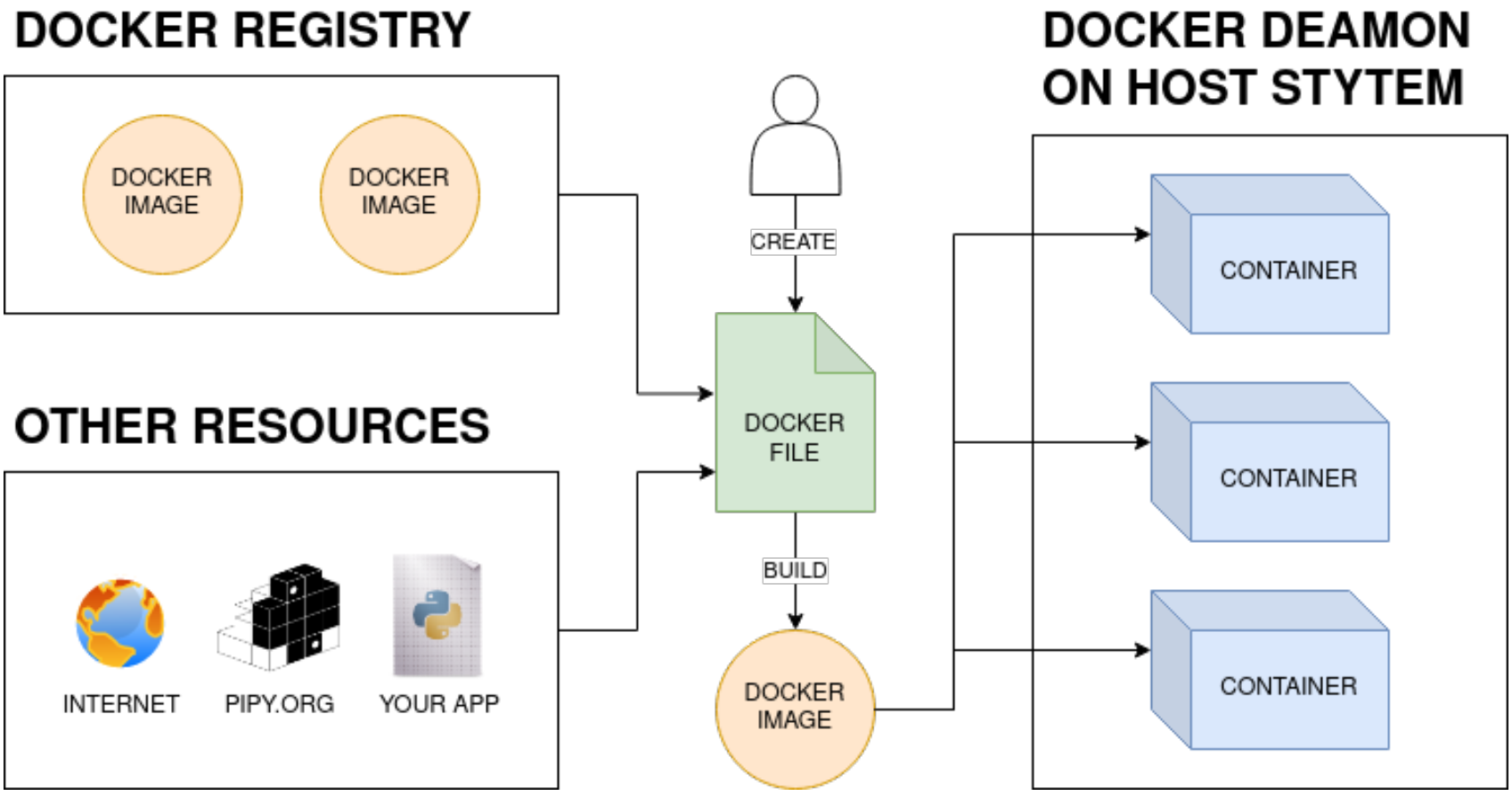
Other sources: Kaggle Datasets, UCI ML repository, Reddit Datasets, BigQuery Datasets, Academic Torrents, AWS Opendata

Environment

Docker

- powerful open source containerization platform
- rapid and continuous deployment of single-purpose machines
- isolates applications in a specified environment
- provides OS isolation, portability and security
- 2021 #1 most wanted and #2 most loved developer tool

https://www.docker.com/


DOCKER REGISTRY
DOCKER IMAGE
DOCKER IMAGE
OTHER RESOURCES
INTERNET
PIPY.ORG
YOUR APP
CREATE
DOCKER FILE
BUILD
DOCKER IMAGE
DOCKER DEAMON ON HOST STYTEM
CONTAINER
CONTAINER
CONTAINER

```
# installing docker.io (package by Debian) on Ubuntu
# check stackoverflow.com/questions/45023363#57678382
# on why you want to do that & what are the trade-offs
sudo apt install docker.io
# for other OS https://docs.docker.com/engine/install/
```

```
# useful commands
docker ps -a # list all containers
docker pull <image_name>  # pull an image from registry
docker images # list all images
docker build <arguments> # build a new image
docker run <arguments> # run a container
docker rm <container id> # remove stopped container
docker rmi <image_name> # remove not used image
docker attach <container id> # connect to container
```

Virtual Environment

- Isolated Python environments
- Addresses the problem of dependencies, versions, and permissions
- Prevents mixing global with project dependencies
- Recreates the whole environment with a few commands

https://virtualenv.pypa.io/

```
sudo pip3 install virtualenv
virtualenv .venv  # create a folder for the env
source .venv/bin/activate  # activate the env
pip install -r requirements.txt  # install all deps
... # do your work
deactivate
```

```
# useful commands
pip --version
pip list # list all installed packages
pip show <packagename> # show package info
pip freeze # output all packages with exact versions
pip install <packagename>
pip install <packagename> --upgrade
pip uninstall <packagename>
```

Jupyter Notebook

- interactive web tool known as a computational notebook
- combines live code and multimedia resources in a single document
- rich interactive output incl. HTML, images, videos, LaTeX
- supports over 40 programming languages

Google Colab

- basically a Jupyter Notebook running in the Google cloud
- write and execute Python in your browser for free
- zero configuration, easy sharing, free access to GPU
- especially well suited to ML, data analysis and education

Versioning

Git

- by far the most used version control system
- open-source project created by Linux Torvalds
- fully distributed, supports offline work
- committing, branching, merging, etc. optimized for performance
- free hosting services are e.g. GitHub.com or GitLab.com
- my favorite guide: rogerdudler.github.io/git-guide/

```
# installing git on Ubuntu
sudo apt install git-all
```

```
# usefull commands
git status # displaying the current status
git init # creating a new repository
git add . # staging latest changes
git commit -m "Commit message" # committing the changes
git push origin master # pushing the commit into origin
git pull # updating local repository
git clone # cloning some repository
```

Weights & Biases

- framework agnostic tool for ML developers
- experiment tracking, dataset versioning, and model management
- hyperparameter optimization, prediction visualization
- exploring and sharing of results from thousands of experiments
- in a centralized cloud dashboard

Presentation

Voila

- turn jupyter notebook into a standalone web application
- have an interactive dashboard for your ML project in mintues
- simple installation with pip
- integrates with Jupyter Notebook or Jupyter Lab
- officially supported by the Jupyter Project

Anvil

- full-stack webapps with nothing but Python
- open-source runtime engine
- compiles your Python app to Javascript
- has a built-in database if you want
- web-based IDE and one click deployment

Compute & Deploy

MyBinder

- environment to run your Jupyter notebooks from you git repo
- provides an easy way for your readers to run your code
- reviewers can reproduce your results without any overhead
- easy to set up and supports Docker
- create a button and place it in your readme.md

Heroku

- fully-managed platform for simple app delivery
- not necessary to manage your own VPS
- supports most languages and deployment with Docker
- simple deployment with one git push command
- free plan (550-1000 hours of runtime per month)
- supports custom domain names on free plan

Publishing

Overleaf

- collaborative cloud-based LaTeX editor
- journal, thesis, presentation templates
- submit your paper directly to a publisher
- seamless collaboration and sharing with co-authors

Google Scholar Metrics

Google Scholar Metrics

- helps authors as they consider where to publish
- journals or conferences ranked by 5y h-index of their papers
- explore the most cited articles and who cited them
- browse by specific research areas of your interest

https://scholar.google.cz/citations?view_op=metrics_intro

Scimago Journal & Country Rank

SJR Scimago Journal & Country Rank

- comprehensive database of journal scientific indicators
- based on data from Scopus and Elsevier
- uses a similar metric as Google PageRank™
- very user friendly interface

https://www.scimagojr.com/

Pavol Harar

pavol.harar.eu

Received an MSc in System Engineering and Informatics and a PhD in Machine Learning from Brno University of Technology. Gained experience in predictive modeling, signal processing, and parallel computing as a member of Brain Diseases Analysis Laboratory and Numerical Harmonic Analysis Group. Currently a postdoc at the Data Science Research Network @UniVie and a visiting postdoc at the Institute of Molecular Pathology in Vienna.